MPLS Working Group
Internet-Draft
Intended status: Informational
Expires: January 13, 2014

A. Atlas
J. Drake
Juniper Networks
S. Giacalone
Thomson Reuters
D. Ward
S. Previdi
C. Filsfils
Cisco Systems

July 12, 2013

Performance-based Path Selection for Explicitly Routed LSPs using TE Metric Extensions
draft-atlas-mpls-te-express-path-03

Abstract

In certain networks, it is critical to consider network performance criteria when selecting the path for an explicitly routed RSVP-TE LSP. Such performance criteria can include latency, jitter, and loss or other indications such as the conformance to link performance objectives and non-RSVP TE traffic load. This specification uses network performance data, such as is advertised via the OSPF and ISIS TE metric extensions (defined outside the scope of this document) to perform such path selections.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on January 13, 2014.

Table of Contents

1. Introduction

In certain networks, such as financial information networks, network performance information is becoming as critical to data path selection as other existing metrics. Network performance information can be obtained via either the TE Metric Extensions in OSPF [I-D.ietf-ospf-te-metric-extensions] or ISIS [I-D.previdi-isis-te-metric-extensions] or via a management system. As with other TE information flooded via OSPF or ISIS, the TE metric extensions have a flooding scope limited to the local area or level. This document describes how to use that information for path selection for explicitly routed LSPs signaled via RSVP-TE [RFC3209]. Methods of optimizing path selection for multiple parameters are generally computationally complex. The methods proposed here are to either make use of either a single metric in path selection, such as minimal path delay, or to make use of another single metric, such as the existing TE metric with additional constraints, such as target link delay bounds and target path delay bounds.

The path selection mechanisms described in this document apply to paths that are fully computed by the head-end of the LSP and then signaled in an ERO where every sub-object is strict. This allows the head-end to consider IGP-distributed performance data without requiring the ability to signal the performance constraints in an object of the RSVP Path message.

When considering performance-based data, it is obvious that there are additional contributors beyond just the links. Clearly end-to-end latency is a combination of router latency, queuing latency, physical link latency and other factors. While traversing a router can cause delay, that can be included in the advertised link delay. As described in [I-D.ietf-ospf-te-metric-extensions] and [I-D.previdi-isis-te-metric-extensions], queuing delay must not be included in the measurements advertised by OSPF or ISIS.

Queuing latency is specifically excluded to insure freedom from oscillations and stability issues that have plagued prior attempts to use delay as a routing metric. If application traffic which follows path based upon latency constraints, the same traffic might be in an Expedited Forwarding Per-Hop-Behavior [RFC3246] with minimal queuing delay or another PHB with potentially very substantial per-hop queuing delay. Only traffic which experiences relatively low congestion, such as Expedited Forwarding traffic, will experience delays very close to the sum of the reported link delays.

This document does not specify how a router determines what values to advertise by the IGP; it does assume that the constraints specified in [I-D.ietf-ospf-te-metric-extensions] and [I-D.previdi-isis-te-metric-extensions] are followed. Additionally, the end-to-end performance that is computed for an LSP path should be built from the individual link data. Any end-to-end characterization used to determine an LSP's performance compliance should be fully reflected in the Traffic Engineering Database so that a CSPF calculation can also determine whether a path under consideration would be in compliance.

1.1. Basic Requirements

The following are the requirements that motivate this solution.

1. Select a TE tunnel's path based upon a combination of existing constraints as well as on link-latency, packet loss, jitter, link performance objectives conformance, and bandwidth consumed by non-RSVP-TE traffic.

2. Ability to define different end-to-end performance requirements for each TE tunnel regardless of common use of resources.

3. Ability to periodically verify that a TE tunnel's current LSP complies with its configured end-to-end performance requirements.

4. Ability to move tunnels, using make-before-break, based upon computed end-to-end performance complying with constraints.

5. Ability to move tunnels away from any link that is violating an underlying link performance objective.

6. Ability to optionally avoid setting up tunnels using any link that is violating a link performance objective, regardless of whether end-to-end performance would still meet requirements.

7. Ability to revert back to the best path after a configurable period.

1.2. Oscillation and Stability Considerations

Past attempts to use unbounded delay or loss as metric sufferred from severe oscillations. The use of performance based data must be such that undampened oscillations are not possible and stability cannot be impacted.

The use of timers is often cited as a cure. Oscillation that is damped by timers is known as "slosh". If advertisement timers are very short relative to the jitter applied to RSVP-TE CSPF timers, then a partial oscillation occurs. If RSVP-TE CSPF timers are short relative to advertisement timers, full oscillation (all traffic moving back and forth) can occur. Even a partial oscillation causes unnecessary reordering which is considered at least minimally disruptive.

Delay variation or jitter is affected by even small traffic levels. At even tiny traffic levels, the probability of a queue occupancy of one can produce a measured jitter proportional to or equal to the packet serialization delay. Very low levels of traffic can increase the probability of queue occupancies of two or three packets enough to further increase the measured jitter. Because jitter measurement is extremely sensitive to even very low traffic levels, any use of jitter is likely to oscillate. There may be legitimate use of a jitter measurement in path computation that can be considered free of oscillation.

Delay measurements that are not sensitive to traffic loads may be safely used in path computation. Delay measurements made at the link layer or measurements made at a queuing priority higher than any significant traffic (such as DSCP CS7 or CS6 [RFC4594], but not CS2 if traffic levels at CS3 and higher or EF and AF can affect the

measurement). Making delay measurements at the same priority as the traffic on affected paths is likely to cause oscillations.

2. Using Performance Data Constraints

2.1. End-to-End Constraints

The per-link performance data available in the IGP [I-D.ietf-ospf-te-metric-extensions] [I-D.previdi-isis-te-metric-extensions] includes: unidirectional link delay, unidirectional delay variation, and link loss. Each (or all) of these parameters can be used to create the path-level link-based parameter.

It is possible to compute a CSPF where the link latency values are used instead of TE metrics, this results in ignoring the TE metrics and causing LSPs to prefer the lowest-latency paths. An alternative to this approach to minimize path latency is an approach to place an upper bound on path latency. An end-to-end latency upper bound merely requires that the path computed be no more than that bound and does not require that it be the minimum latency path. This upper bound can be used as a constraint in CSPF to prevent exploring links that would create a path over the end-to-end latency bound. Both approaches are valid.

Basic pseudo-code to further explain this pruning is given below. The change is to decide whether to explore a link based on whether the resulting path would violate the specified latency-bound and to accumulate the latency used along the path.

```
  CSPF_with_bound(root, latency_bound)
    root.distance = 0
    root.latency = 0
    fibheap_insert(root, root.distance)
    while (fibheap_not_empty())
       next_rtr = fibheap_pop
       for each link next_link attached to next_rtr
          if (next_rtr.latency + next_link.latency) < latency_bound
              explore_link(next_rtr, next_link)

  Explore_Link(next_rtr, next_link)
    if ((next_rtr.distance + next_link.metric) <
        next_link->neighbor.distance)
       next_link->neighbor.distance = next_rtr.distance +
                                      next_link.metric
       next_link->neighbor.latency = next_rtr.latency +
                                     next_link.latency
```

```
    if ((next_rtr.distance + next_link.metric) ==
        next_link->neighbor.distance)
       if ((next_rtr.latency + next_link.latency) <
            next_link->neighbor.latency)
           next_link->neighbor.latency = next_rtr.latency +
                                         next_link.latency
```

Partial Pseudo-Code for latency-pruned SPF

An end-to-end bound on delay variation can be used similarly as a constraint in the CSPF on what links to explore where the path's delay variation is the sum of the used links' delay variations.

For link loss, the path loss is not the sum of the used links' losses. Instead, the path loss percentage is 100 - (100 - loss_L1)*(100 - loss_L2)*...*(100 - loss_Ln), where the links along the path are L1 to Ln. The end-to-end link loss bound, computed in this fashion, can also be used as a constraint in the CSPF on what links to explore.

2.2. Link Constraints

In addition to selecting paths that conform to a bound on performance data, it is also useful to avoid using links that do not meet a necessary constraint. Naturally, if such a parameter were a known fixed value, then resource attribute flags could be used to express this behavior. However, when the parameter associated with a link may vary dynamically, there is not currently a configuration-time mechanism to enforce such behavior. An example of this is described in Section 2.3, where links may move in and out of conformance for link performance objectives with regards to latency, delay variation, and link loss.

When doing path selection for TE tunnels, it has not been possible to know how much actual bandwidth is available that includes the bandwidth used by non-RSVP-TE traffic. In [I-D.ietf-ospf-te-metric-extensions] [I-D.previdi-isis-te-metric-extensions], the Unidirectional Available Bandwidth is advertised as is the Residual Bandwidth. When computing the path for a TE tunnel, only links with at least a configurable amount of Unidirectional Available Bandwidth might be permitted.

Similarly, only links whose loss is under a configurable value might be acceptable. For these constraints, each link can be tested against the constraint and only explored in the CSPF if the link passes. In essence, a link that fails the constraint test is treated as if it contained a resource attribute in the exclude-any filter.

2.3. Links out of compliance with Link Performance Objectives

Link conformance to a link performance objective can change as a result of rerouting at lower layers. This could be due to optical regrooming or simply rerouting of a FA-LSP. When this occurs, there are two questions to be asked:

a. Should the link be trusted and used for the setup of new LSPs?

b. Should LSPs using this link automatically be moved to a secondary path?

2.3.1. Use of Anomalous Links for New Paths

If the answer to (a) is no for link latency performance objectives, then any link which has the Anomalous bit set in the Unidirectional Link Delay sub-TLV[I-D.ietf-ospf-te-metric-extensions] [I-D.previdi-isis-te-metric-extensions] should be removed from the topology before a CSPF calculation is used to compute a new path. In essence, the link should be treated exactly as if it fails the exclude-any resource attributes filter.[RFC3209].

Similarly, if the answer to (a) is no for link loss performance objectives, then any link which has the Anomalous bit set in the Link Los sub-TLV should be treated as if it fails the exclude-any resource attributes filter. If the answer to (a) is no for link jitter performance objectives, then any link that has the Anomalous bit set in the Unidirectional Delay Variation sub-TLV[I-D.previdi-isis-te-metric-extensions] should be treated as if it fails the exclude-any resource attributes filter.

2.3.2. Links entering the Anomalous State

When a link enters the Anomalous state with respect to a parameter, this is an indication that LSPs using that link might also no longer be in compliance with their performance bounds. It can also be considered an indication that something is changing that link and so it might no longer be trustworthy to carry performance-critical traffic. Naturally, which performance criteria are important for a particular LSP is dependent upon the LSP's configuration and thus the compliance of a link with respect to a particular link performance objective is indicated per performance criterion.

At the ingress of a TE tunnel, a TE tunnel may be configured to be sensitive to the Anomalous state of links in reference to latency, delay variation, and/or loss. Additionally, such a TE tunnel may be configured to either verify continued compliance, to switch immediately to a standby LSP, or to move to a different path.

When a sub-TLV is received with the Anomalous bit set when previously it was clear, the list of interested TE tunnels must be scanned. Each such TE tunnel should either have its continued compliance verified, be switched to a hot standby, or do a make-before-break to a secondary path.

It is not sufficient to just look at the Anomalous bit in order to determine when TE tunnels must have their compliance verified. When changing to set, the Anomalous bit merely provides a hint that interested TE tunnels should have their continued compliance verified.

2.3.3. Links leaving the Anomalous State

When a link leaves the Anomalous state with respect to a parameter, this can serve as an indication that those TE tunnels, whose LSPs were changed due to administrative policy when the link entered the Anomalous state, may want to reoptimize to a better path. The hint provided by the Anomalous state change may help optimize when to recompute for a better path.

3. IANA Considerations

This document includes no request to IANA.

4. Security Considerations

This document is not currently believed to introduce new security concerns.

5. Acknowledgements

The authors would like to thank Curtis Villamizar for his extensive detailed comments and suggested text in the Section 1 and Section 1.2. The authors would also like to thank Xiaohu Xu and Sriganesh Kini for their review.

6. References

6.1. Normative References

[I-D.ietf-ospf-te-metric-extensions]
Giacalone, S., Ward, D., Drake, J., Atlas, A., and S. Previdi, "OSPF Traffic Engineering (TE) Metric Extensions", draft-ietf-ospf-te-metric-extensions-04 (work in progress), June 2013.

[I-D.previdi-isis-te-metric-extensions]

Previdi, S., Giacalone, S., Ward, D., Drake, J., Atlas, A., and C. Filsfils, "IS-IS Traffic Engineering (TE) Metric Extensions", draft-previdi-isis-te-metric-extensions-03 (work in progress), February 2013.

[RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.

6.2. Informative References

[RFC3246] Davie, B., Charny, A., Bennet, J., Benson, K., Le Boudec, J., Courtney, W., Davari, S., Firoiu, V., and D. Stiliadis, "An Expedited Forwarding PHB (Per-Hop Behavior)", RFC 3246, March 2002.

[RFC4594] Babiarz, J., Chan, K., and F. Baker, "Configuration Guidelines for DiffServ Service Classes", RFC 4594, August 2006.

Authors' Addresses

Alia Atlas
Juniper Networks
10 Technology Park Drive
Westford, MA 01886
USA

Email: akatlas@juniper.net

John Drake
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
USA

Email: jdrake@juniper.net

Spencer Giacalone
Thomson Reuters
195 Broadway
New York, NY 10007
USA

Email: Spencer.giacalone@thomsonreuters.com

Dave Ward
Cisco Systems
170 West Tasman Dr.
San Jose, CA 95134
USA

Email: dward@cisco.com

Stefano Previdi
Cisco Systems
Via Del Serafico 200
Rome 00142
Italy

Email: sprevidi@cisco.com

Clarence Filsfils
Cisco Systems
Brussels
Belgium

Email: cfilsfil@cisco.com